QUEUEING THEORY AND MODELING

Linda Green
Graduate School of Business,Columbia University,New York, New York 10027

Abstract: Many organizations, such as banks, airlines, telecommunications companies, and police departments, routinely use queueing models to help manage and allocate resources in order to respond to demands in a timely and cost-efficient fashion. Though queueing analysis has been used in hospitals and other healthcare settings, its use in this sector is not widespread. Yet, given the pervasiveness of delays in healthcare and the fact that many healthcare facilities are trying to meet increasing demands with tightly constrained resources, queueing models can be very useful in developing more effective policies for allocating and managing resources in healthcare facilities. Queueing analysis is also a useful tool for estimating capacity requirements and managing demand for any system in which the timing of service needs is random. This chapter describes basic queueing theory and models as well as some simple modifications and extensions that are particularly useful in the healthcare setting, and gives examples of their use. The critical issue of data requirements is also discussed as well as model choice, model-building and the interpretation and use of results.

Key words: Queueing, capacity management, staffing, hospitals

Introduction

Why are queue models helpful in healthcare?

Healthcare is riddled with delays. Almost all of us have waited for days or weeks to get an appointment with a physician or schedule a procedure, and upon arrival we wait some more until being seen. In hospitals, it is not unusual to find patients waiting for beds in hallways, and delays for surgery or diagnostic tests are common.

Delays are the result of a disparity between demand for a service and the capacity available to meet that demand. Usually this mismatch is temporary and due to natural variability in the timing of demands and in the duration of time needed to provide service. A simple example would be a healthcare clinic where patients walk in without appointments in an unpredictable fashion and require anything from a flu shot to the setting of a broken limb. This variability and the interaction between the arrival and service processes make the dynamics of service systems very complex. Consequently, it’s impossible to predict levels of congestion or to determine how much capacity is needed to achieve some desired level of performance without the help of a queueing model.

Queueing theory was developed by A.K. Erlang in 1904 to help determine the capacity requirements of the Danish telephone system (see Brockmeyer et al. 1948). It has since been applied to a large range of service industries including banks, airlines, and telephone call centers (e.g. Brewton 1989, Stern and Hersh 1980, Holloran and Byrne 1986, Brusco et al 1995, and Brigandi et al 1994) as well as emergency systems such as police patrol, fire and ambulances (e.g. Larson 1972, Kolesar et al 1975, Chelst and Barlach 1981, Green and Kolesar 1984, Taylor and Huxley 1989). It has also been applied in various healthcare settings as we will discuss later in this chapter. Queueing models can be very useful in identifying appropriate levels of staff, equipment, and beds as well as in making decisions about resource allocation and the design of new services.

Unlike simulation methodologies, queueing models require very little data and result in relatively simple formulae for predicting various performance measures such as mean delay or probability of waiting more than a given amount of time before being served. This means that they are easier and cheaper to develop and use. In addition, since they are extremely fast to run, they provide a simple way to perform "what-if" analyses, identify tradeoffs and find attractive solutions rather than just estimating performance for a given scenario.

Timely access has been identified as one of the key elements of healthcare quality (Institute of Medicine 2001) and consequently, decreasing delays has become a focus in many healthcare institutions. Given the financial constraints that exist in many of these organizations, queueing analysis can be an extremely valuable tool in utilizing resources in the most cost-effective way to reduce delays. The primary goal of this chapter is to provide a basic understanding of queueing theory and some of the specific queueing models that can be helpful in designing and managing healthcare delivery systems. For more detail on specific models that are commonly used, a textbook on queueing theory such as Hall (1991) is recommended.

Before discussing past and potential uses of queueing models in healthcare, it’s important to first understand some queueing theory fundamentals.

Queueing Fundamentals

A basic queueing system is a service system where “customers” arrive to a bank of “servers” and require some service from one of them. It’s important to understand that a “customer” is whatever entity is waiting for service and does not have to be a person. For example, in a “back-office” situation such as the reading of radiologic images, the “customers” might be the images waiting to be read. Similarly, a “server” is the person or thing that provides the service. So when analyzing delays for patients in the emergency department (ED) awaiting admission to the hospital, the relevant servers would be inpatient beds.

If all servers are busy upon a customer’s arrival, they must join a queue. Though queues are often physical lines of people or things, they can also be invisible as with telephone calls waiting on hold. The rule that determines the order in which queued customers are served is called the queue *discipline*. The most common discipline is the familiar first-come, first-served (FCFS) rule, but other disciplines are often used to increase efficiency or reduce the delay for more time-sensitive customers. For example, in an ED, the triage system is an example of a *priority* queue discipline. Priority disciplines may be preemptive or non-preemptive, depending upon whether a service in progress can be interrupted when a customer with a higher priority arrives. In most queueing models, the assumption is made that there is no limit on the number of customers that can be waiting for service, i.e. there is an *infinite waiting room*. This may be a good assumption when customers do not physically join a queue, as in a telephone call center, or when the physical space where customers wait is large compared to the number of customers who are usually waiting for service. Even if there is no capacity limit on waiting room, in some cases new arrivals who see a long queue may “balk” and not join the queue. This might happen in a walk-in clinic. A similar behavior that is incorporated in some queueing systems is “reneging” or "abandonment" which occurs when customers grow inpatient and leave the queue before being served. An example of this behavior is found in some EDs where the patients who renege are often referred to as “left without being seen”.

Finally, queues may be organized in various ways. In most cases, we will consider a *single line* that feeds into all servers. But sometimes each server has his/her own queue as may be the case for a primary care office in which patients have their own physician. This design is usually referred to as queues in *parallel*. In other situations, we may want to consider a *network* design in which customers receive service from different types of servers in a sequential manner. For example, a surgical inpatient requires an operating room (OR), then a bed in the recovery unit, followed by a bed in a surgical intensive care unit (ICU), and/or other part of the hospital. However, it might still make sense to analyze one specific single queue in these situations to determine the capacity requirements of a single type of resource, particularly if there is reason to believe that the resource is a bottleneck.

A queueing model is a mathematical description of a queuing system which makes some specific assumptions about the probabilistic nature of the arrival and service processes, the number and type of servers, and the queue discipline and organization. There are countless possible variations, but some queueing models are more widely used and we will focus on these in this chapter. For these models, as well as many others, there are formulae available that enable the fast calculation of various performance measures that can be used to help design a new service system or improve an existing one.

BASIC QUEUEING PRINCIPLES AND MODELS

Most of queueing theory deals with system performance in *steady-state.* That is, most queueing models assume that the system has been operating with the same arrival rate, average service time and other characteristics for a sufficiently long time that the probabilistic behavior of performance measures such as queue length and customer delay is independent of when the system is observed. Clearly, there are many service systems, including health care systems, for which there are time-of-day, day-of-week or seasonality affects. In this section, we will assume that we are looking at systems in steady-state and in subsequent sections, we will discuss how to deal with systems that have some time-varying characteristics.

Delays, Utilization and System Size

In queueing theory, utilization, defined as the average number of busy servers divided by the total number of servers times 100, is an important measure. From a managerial perspective, utilization is often seen as a measure of productivity and therefore it is considered desirable for it to be high. For example, in hospital bed planning, utilization is called occupancy level and historically, an average hospital occupancy level of 85 percent has been used as the minimum level for the states to make a determination under Certificate of Need (CON) regulations that more beds might be needed (see Brecher and Speizio 1995). Since the actual average occupancy level for nonprofit hospitals has recently been under 70 percent, there has been a widely held perception in the health care community that there are too many hospital beds. Largely because of this perception, the number of hospital beds has decreased almost 25 percent in the last 20 years.

But determining bed capacity based on occupancy levels can result in very long waiting times for beds (Green 2003). In all queueing systems, the higher the average utilization level, the longer the wait times. However, it is important to note that this relationship is nonlinear. This is illustrated in Figure 1 which shows the fundamental relationship between delays and utilization for a queueing system. There are three critical observations we can make from this figure. First, as average utilization (e.g. occupancy level) increases, average delays increase at an increasing rate. Second, there is an "elbow" in the curve after which the average delay increases more dramatically in response to even small increases in utilization. Finally, the average delay approaches infinity as utilization approaches one. (It's important to note that this is assuming that there is no constraint on how long the queue can get and that customers continue to join and remain in the queue.)

The exact location of the elbow in the curve depends upon two critical characteristics of the system: variability and size. Variability generally exists in both the time between arrivals and the duration of service times and is usually measured by the ratio of the standard deviation to the mean, called the coefficient of variation (*CV*). The higher the degree of variability in the system, the more to the left the elbow will be so that delays will be worse for the same utilization level. System size is defined as the ratio of the average demand over the average service time, which is a determinant of the number of servers needed. The larger the system, the closer the elbow will be to 100%, so that delays will be shorter for the same utilization level.

These basic queueing principles have several important implications for planning or evaluating capacity in a service system. First, the average total capacity, defined as the number of servers times the rate at which each server can serve customers, must be strictly greater than the average demand. In other words, unless average utilization is strictly *less than* 100%, the system will be "unstable" and the queue will continue to grow. Though this fact may appear counter-intuitive on the surface, it has been well known by operations professionals for decades. So if an emergency room has 10 patients arriving per hour on average and each healthcare provider (physician or physician assistant) can treat 2 patients per hour on average, a minimum of 6 providers is needed. (Of course, in many contexts, if arrivals see a long queue they may not join it or they may renege after waiting a long time. If so, it may be possible to have stability even if the average demand exceeds the average capacity.) Second, the smaller the system, the longer the delays will be for a given utilization level. In other words, queueing systems have economies of scale so that, for example, larger hospitals can operate at higher utilization levels than smaller ones yet maintain similar levels of congestion and delays. Finally, the greater the variability in the service time (e.g. length-of-stay), the longer the delays at any given utilization level. So a clinic or physician office that specializes in e.g. vision testing or mammography, will experience shorter patient waits than a university based clinic of the same size and with the same provider utilization that treats a broad variety of illnesses and injuries. These properties will be more specifically illustrated when we discuss applications of queueing models.

Some simple but useful queueing models

The Poisson process

In specifying a queueing model, we must make assumptions about the probabilistic nature of the arrival and service processes. The most common assumption to make about arrivals is that they follow a *Poisson* process. The name comes from the fact that the number of arrivals in any given time period has a Poisson distribution. So if *N(t)* is the number of arrivals during a time period of duration t and N(t) has a Poisson distribution,

$$\text{Probability } \{N(t) = \text{n}\} = \text{e}^{-\lambda t}\,(\lambda t)^{n} / \text{n}!$$

where λ is called the *rate* and is the expected number of arrivals per unit time. For example, if $\lambda = 10$ customers per hour, then the expected number of arrivals in any 60 minute interval is 10 and the expected number to arrive in a 15 minute interval is 2.5. Notice that these are averages so that λ need not have an integer value. Another way to characterize the Poisson process is that the time between consecutive arrivals, called the interarrival time, has an *exponential* distribution. So if *IA* is the interarrival time of a Poisson process with rate λ,

$$\text{Probability } \{IA \le \text{t}\} = 1 - \text{e}^{-\lambda t}$$

and 1/λ is the average time between arrivals.

An important property of the exponential distribution is that it is "memoryless". This means that the time of the next arrival is independent of when the last arrival occurred. This property also leads to the fact that if the arrival process is Poisson, the number of arrivals in any given time interval is independent of the number in any other non-overlapping time interval. Conversely, it can be shown analytically that if customers arrive independently from one another, the arrival process is a Poisson process. For this reason, the Poisson process is considered the most "random" arrival process.

In determining whether the Poisson process is a reasonable model for arrivals in a specific service system, it is useful to consider its three defining properties:

1. Customers arrive one at a time.
2. The probability that a customer arrives at any time is independent of when other customers arrived.
3. The probability that a customer arrives at a given time is independent of the time.

In most contexts, customers generally do arrive one at a time. Though there may be events, such as a major accident, that trigger multiple simultaneous arrivals, this is likely to be an exceptional circumstance which will not significantly affect the effectiveness of this modeling assumption. Intuitively, the second property is also often a reasonable assumption. For example, in an emergency room, where the population of potential patients is very large, it is unlikely that someone arriving with a broken arm has anything to do with someone else's injury or illness, or that the fact that the number of patients who arrived between 9am and 10am was four provides any information about the number of patients that are likely to arrive between 10am and 11am. Again, there may be occasional exceptions, such as a flu outbreak, which violate this assumption, but in the aggregate, it's likely to be reasonable. However, the third property may be more suspect. More typically, the average arrival rate varies over the day so that, e.g., it is more likely for an arrival to occur in the morning than in the middle of the night. Certain days of the week may be busier than others as well. However, we may be able to use the standard Poisson process as a good model for a shorter interval of time over which the arrival rate is fairly constant. We will discuss this in more detail in a subsequent section.

So the assumption of a Poisson process will generally be a good one when the three properties above are a reasonable description of the service system in question. However, it is possible to perform more rigorous tests to determine if it is a good fit. The simplest tests are based on the relationship of the standard deviation to the mean of the two distributions involved in the Poisson process. Since the variance (square of the standard deviation) of the Poisson distribution is equal to its mean, we can examine the number of arrivals in each fixed interval of time, (e.g. 30 minutes) and determine whether the ratio of the mean to the variance is close to one. Alternatively, since the exponential distribution is characterized by its standard deviation being equal to its mean, we can look at the interarrival times and compute the ratio of the standard deviation to the mean to see if it's close to one. Hall (1991) describes goodness of fit tests in greater detail.

Many real arrival and demand processes have been empirically shown to be very well approximated by a Poisson process. Among these are demands for emergency services such as police, fire and ambulance, arrivals to banks and other retail establishments, and arrivals of telephone calls to customer service call centers. Because of its prevalence and its assumption of independent arrivals, the Poisson process is the most commonly used arrival process in modeling service systems. It is also a convenient assumption to make in terms of data collection since it is characterized by a single parameter – its rate λ. In healthcare, the Poisson process has been verified to be a good representation of unscheduled arrivals to various parts of the hospital including ICUs, obstetrics units and EDs (Young 1965, Kim et al 1999, Green et al 2005).

The M/M/s *model*

The most commonly used queueing model is the *M/M/s* or *Erlang delay* model. This model assumes a single queue with unlimited waiting room that feeds into *s* identical servers. Customers arrive according to a Poisson process with a constant rate, and the service duration (e.g. LOS or provider time associated with a patient) has an exponential distribution. (These two assumptions are often called Markovian, hence the use of the two "M's" in the notation used for the model.)

One advantage of using the *M/M/s* model is that it requires only three parameters and so it can be used to obtain performance estimates with very little data. Given an average arrival rate, λ, an average service duration, 1/μ, and the number of servers, *s*, easy-to-compute formulae are available to obtain performance measures such as the probability that an arrival will experience a positive delay, p_D, or the average delay, W_q:

$$p_D = 1 - \sum_{n=0}^{s-1} p_n \tag{1}$$

$$W_q = p_D / [(1 - \rho)\, s\mu] \tag{2}$$

for

$$\rho = \lambda / s\mu \tag{3}$$

and

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\,\mu^n} p_0 & (1 \le n \le s) \\ \dfrac{\lambda^n}{s^{n-s} s!\,\mu^n} p_0 & (n \ge s) \end{cases} \tag{4}$$

where

$$p_0 = \left[\sum_{n=0}^{s-1} \frac{(\rho s)^n}{n!} + \frac{\rho^s s^{s+1}}{s!(s-\rho s)} \right]^{-1} \qquad \rho < 1 \tag{5}$$

Note that ρ is the average utilization for this queueing system and the equation is only valid when the utilization is strictly less than one. Also note that average delay increases as utilization approaches one. These quantitative observations support the discussion of utilization and delays in the previous section.

Many other measures of performance can be calculated as well and many of the formulae for both the *M/M/s* and other common queueing models are available in software packages or are easily programmable on spreadsheets. One common performance constraint is often referred to as the *service level* – a requirement that *x*% of customers start service within *y* time units. For example, many customer call centers have a target service level that 85% of calls be answered within 20 seconds. The delay is always measured from the time of the demand for service (e.g. patient registered in the ED) to the time at which service begins (e.g. a provider is available to treat that patient). It's important to note that the model's delay predictions pertain only to waiting times due to the unavailability of the server.

Some useful extensions of the M/M/s model

There are several variations on the basic *M/M/s* queueing model. One important one for many healthcare organizations is the *M/M/s* with priorities. While the fundamental model assumes that customers are indistinguishable and are served FCFS, the priority model assumes that customers have differing time-sensitivities and are allocated to two or more service classes *i* = 1,2,…*N*, and that customers are served in priority order with 1 being the highest priority and *N* the lowest. Within any given class, customers are served FCFS. But when there is a queue and a server becomes available, a customer belonging to class i > 1 will be served only if there are no waiting customers of class 1,…,*i*-1. A priority queueing model would be appropriate if a facility is interested in identifying the capacity needed to assure a targeted service level for the highest priority customers. For example, in an ED, while many arriving patients would not incur any particular harm if they had to wait more than an hour to be seen by a physician, some fraction, who are emergent or urgent, need a physician's care sooner to prevent serious clinical consequences. In this case, a priority queueing model could be used to answer a question like: How many physicians are needed to assure that 90% of emergent and urgent patients will be seen by a physician within 45 minutes?

There are two types of priority queueing disciplines: preemptive and non-preemptive. In the preemptive model, if a higher priority customer arrives when all servers are busy and a lower priority customer is being served, the lower priority customer's service will be interrupted (preempted) so that the higher priority customer can begin service immediately. The preempted customer must then wait for another server to become free to resume service. In the non-preemptive model, new arrivals cannot preempt customers already in service. While priority queueing models are usually either purely preemptive or non-preemptive, it is possible to model a service system that has both preemptive and non-preemptive customer classes. This might be appropriate for a hospital ED where the normal triage system which classifies patients as emergent, urgent or non-urgent is generally assumed to be non-preemptive, but a preemptive discipline is used for certain urgent patients whose conditions are extremely time-sensitive, such as stroke victims. In addition to the usual input parameters for the *M/M/s* model, priority models also require data on the fraction of customers in each of the priority classes.

Another common variant of the *M/M/s* model assumes a finite capacity $K \geq s$ and is notated as *M/M/s/K*. In this model, if a customer arrives when there are K customers already in the system (being served and waiting), the customer cannot join the queue and must leave. A common application of this would be a telephone trunk line feeding into a call center. Such a system has a finite number of spaces for calls being served or on hold and when a new call comes in and all the spaces are already filled , the new arrival hears a busy signal and hangs up. A similar phenomenon might occur in a walk-in health clinic which has a waiting room with a fixed number of seats. Though some patients may choose to wait even if there is no seat available upon arrival, many patients may leave and try to return at a less busy time. Customers who are "blocked" from joining the queue are called "lost" and may show up again or never return. In these types of systems, queueing analysis might be used to help determine how large the waiting or holding area should be so that the number of customers who are blocked is kept to an acceptably low level.

A specific special case of these finite capacity models is the one where $K = s$ so that there is no waiting room for those who arrive when all servers are busy. These are called pure "loss" models and they are often used to analyze service systems in which it is considered either impractical or very undesirable to have any customers wait to begin service. For example, Kaplan, Sprung and Shmueli (2003) used a loss model to analyze the impact of various admissions policies to hospital intensive care units.

The M/G/1 *and* G/G/s *models*

An important characteristic of the exponential distribution used in the *M/M/s* is that the standard distribution equals the mean and so the CV of the service time equals one. If the actual CV of service is a bit less than or greater than one, the *M/M/s* will still give good estimates of delay. However, if the CV is substantially different than one, the *M/M/s* may significantly underestimate or overestimate actual delays. (Recall that if variability is lower, the model will overestimate delays while the converse is true if variability is greater.) In this case, if the arrival process is Poisson, and there is only one server, the average delay can still be calculated for any service distribution through use of the following formula for what is known as the *M/G/1* system:

$$W_q = [\lambda\rho / (1 - \rho)] [(1 + CV^2(S)) / 2] \qquad (6)$$

where $CV^2(S)$ is the square of the coefficient of variation of the service time. Clearly, this formula requires knowledge of the standard deviation of the service time in addition to the mean in order to compute $CV^2(S)$. This formula also illustrates the impact of variability on delays. Notice that, as mentioned previously, as the coefficient of variation of the service time increases, so does the average delay.

Though there are no exact formula for non-Markovian multi-server queues, there are some good, simple approximations. One such approximation (Allen 1978) is given by:

$$W_q = W_{q,M/M/s} \ [CV^2(A) + CV^2(S)] / 2 \qquad (7)$$

where $CV^2(A)$ is the square of the coefficient of variation of the arrival time and $W_{q,M/M/s}$ is the expected delay for an *M/M/s* system, eq. (2). So this formula requires the standard deviation of the interarrival time as well and again demonstrates that more variability results in longer delays.

ANALYSES OF FIXED CAPACITY: HOW MANY HOSPITAL BEDS?

Many resources in health care facilities have a fixed capacity over a long period of time. These are usually “things” rather than people: beds, operating rooms, imaging machines, etc. Queueing models are not always appropriate for analyzing such resources. In particular, if the patients for a resource are scheduled into fixed time slots, there is little or no likelihood of congestion unless patients routinely come late or the time slots are not large enough to accommodate most patients. An example of this would be a magnetic resonance imaging (MRI) facility which is only used by scheduled outpatients.

However, the difficulty of managing many healthcare facilities is that the demand for resources is unscheduled and hence random, yet timely care is important. This is the case for many parts of a hospital that deal primarily with non-elective admissions. In these cases, queueing models can be very helpful in identifying long-term capacity needs.

Applying the **M/M/s** *model*

To illustrate the use of a queueing model for evaluating capacity, consider an obstetrics unit. Since it is generally operated independently of other services, its capacity needs, e.g. number of postpartum beds, can be determined without regard to other parts of the hospital. It is also one for which the use of a standard *M/M/s* queueing model is quite good. Most obstetrics patients are unscheduled and the assumption of Poisson arrivals has been shown to be a good one in studies of unscheduled hospital admissions (Young 1965). In addition, the *CV* of length of stay is typically very close to 1.0 (Green and Nguyen 2001) satisfying the service time assumption of the *M/M/s* model.

A queueing model may be used either descriptively or prescriptively. As an example of the descriptive case, we can take the current operating characteristics of a given obstetrics unit: arrival rate, average LOS, and number of beds; and use these in equation (1) to determine the probability that an arriving patient will not find a bed available. Let’s assume that Big City Hospital’s obstetrics unit has an average arrival rate of $\lambda = 14.8$ patients per day, an average LOS of $1/\mu = 2.9$ days, and $s = 56$ beds. Then the *M/M/s* formula for probability of delay (1) produces an estimate of approximately 4%. To use the *M/M/s* prescriptively to find the minimum number of beds needed to attain a target probability of delay, we can enter equation (1) in a spreadsheet and produce a table of results for a broad range of bed capacities to find the one that best meets the desired target. Table 1 is a partial table of results for our example obstetrics unit.

Though there is no standard delay target, Schneider (1981) suggested that given their emergent status, the probability of delay for an obstetrics bed should not exceed 1%. Applying this criterion, Table 1 indicates that this unit has at least 60 beds. Table 1 also shows the utilization level for each choice of servers and that at 60 beds, this level is 71.5%. This is what hospitals call the average occupancy level and it is well below the 85% level that many hospitals and healthcare policy officials consider the minimum target level. It is also below the maximum level of 75% recommended by the American College of Obstetrics and Gynecology (ACOG) to assure timely access to a bed (Freeman and Poland 1992). So does this example show that as long as an obstetrics unit operates below this ACOG occupancy level of 75%, the fraction of patients who will be delayed in getting a bed will be very low?

The problem with using target occupancy levels

Hospital capacity decisions traditionally have been made, both at the government and institutional levels, based on target occupancy levels - the average percentage of occupied beds. Historically, the most commonly used occupancy target has been 85%. Estimates of the number of "excess" beds in the United States, as well as in individual states and communities, usually have been based on this "optimal" occupancy figure (Brecher and Speizio 1995, p.55). In addition, low occupancy levels are often viewed as indicative of operational inefficiency and potential financial problems. So hospital administrators generally view higher occupancy levels as desirable. However, as we saw previously in this chapter, higher occupancy levels result in longer delays and so basing capacity on target occupancy levels can lead to undesirable levels of access for patients.

In Green (2002), the basic *M/M/s* model was used to demonstrate the implications of using target occupancy levels to determine capacity in both obstetrics and ICU units in New York State. Figure 1 from that paper (shown here as Figure 2) shows the distribution of average occupancy rates for 148 obstetrics units in New York State for 1997. These data, representing nearly all obstetrics units in New York, were obtained from Institutional Cost Reports (ICRs), and unlike most other published data, reflect staffed beds rather than certified beds. The graph shows that many maternity units had low average occupancy levels with the overall average occupancy level for the study hospitals was only 60%, which, based on the ACOG standard, would imply significant excess capacity. Applying this 75% standard to the 1997 data, 117 of the 148 New York state hospitals had "excess" beds, while 27 had insufficient beds.

However, if one considers a bed delay target as a more appropriate measure of capacity needs, the conclusions can be quite different. Now the number of beds in each unit becomes a major factor since, for a given occupancy level, delays increase as unit size decreases. While obstetrics units usually are not the smallest units in a hospital, there are many small hospitals, particularly in rural areas, and the units in these facilities may contain only five to 10 beds. Of the New York state hospitals considered here, more than 50% had maternity units with 25 or fewer beds.

In the *M/M/s* model, probability of delay is a function of only two parameters: s and ρ, which in our context is the number of beds and occupancy level. Each of the three curves shown in Figure 3 represents a specific probability of delay as a function of these two variables as generated by equation (1). Thus, using the unit size and occupancy level reported on the ICR report for a given maternity unit, we can determine from this figure if the probability of delay meets or exceeds any one of these targets. For example, if a maternity unit has 15 beds and an occupancy level of 45%, it would fall below all three curves and hence have a probability of delay less than .01 or 1%, meeting all three targets.

Doing this for every hospital in the database, 30 hospitals had insufficient capacity based on even the most slack delay target of 10%. (It is interesting to note that two of the hospitals that would be considered over utilized under the 75% occupancy standard had sufficient capacity under this delay standard.) Tightening the probability of delay target to 5%, yields 48 obstetrics units that do not meet this standard. And adopting a maximum probability of delay of 1% as was suggested in the only publication identified as containing a delay standard for obstetrics beds (Schneider 1981), results in 59, or 40%, of all New York state maternity units with insufficient capacity.

How many hospitals in New York State had maternity units large enough to achieve the ACOG-suggested 75% occupancy level and also meet a specified probability of delay standard? Using Figure 3, we see that for a 10% target, an obstetrics unit would need to have at least 28 beds, a size that exists in only 40% of the state hospitals. For a 5% standard, the minimum number of beds needed is 41, a size achieved in only 14% of the hospitals; for a 1% standard, at least 67 beds are needed, leaving only three of the 148 or 2% of the hospitals of sufficient size.

Choosing a delay standard

As the previous analysis illustrates, the number of required beds can change substantially depending upon what level of delay is considered tolerable. There is no single right choice and in choosing a delay standard, several factors are relevant.

First, what is the expected delay of those patients who experience a delay? This performance measure can be easily calculated once both the probability of delay (equation 1) and the average or mean delay (equation 2) are known. Specifically,

$$\text{Expected delay of delayed customers} = W_q \,/\, p_D \qquad (8)$$

So returning to our obstetrics example above, Table 1 shows that the average delay is .008 days (note that since the input was expressed in days, so is the output) which multiplying by 24 gives us .19 hours. So dividing this by the probability of delay of .04 results in an expected delay for delayed patients of about 4.75 hours. This may indicate that the probability of delay standard should be lower. This, of course, should be considered in light of what this level of congestion means for the particular hospital.

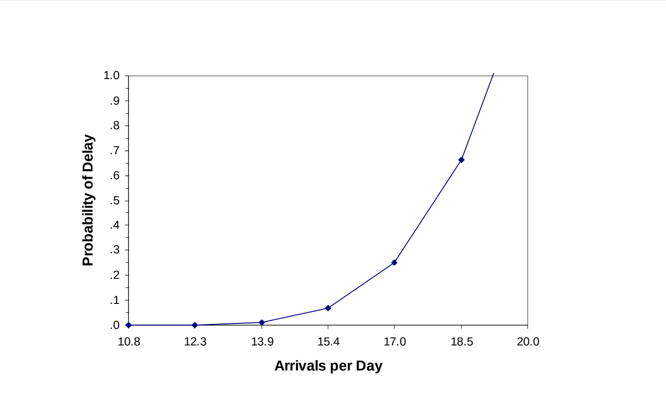
What are the possible consequences of congestion? In the obstetrics case, while patients in some hospitals remain in the same bed through labor, delivery, recovery, and postpartum, in most maternity units, there are separate areas for some or all of these stages of birth. Therefore, a delay for an obstetrics bed often means that a postpartum patient will remain in a recovery bed longer than necessary. This, of course, may cause a backup in the labor and delivery areas so that newly arriving patients may have to wait on gurneys in hallways or in the emergency room. Some hospitals have overflow beds in a nearby unit that is opened (staffed) when all regular beds are full. (This is likely the case for the five hospitals that reported average occupancy levels exceeding 100%.) While these effects of congestion likely pose no medical threat for most patients who experience normal births, there could be adverse clinical consequences in cases in which there are complications. In particular, whether patients are placed in hallways or overflow units, the nursing staff is likely to be severely strained, thereby limiting the quantity and quality of personal attention. Even if a hospital is able to obtain additional staffing, it is usually by using agency nurses who are more expensive and not as familiar with the physical or operating environment, thereby jeopardizing quality of patient care. In addition, telemetry devices, such as fetal monitors that are usually in labor and delivery rooms, may be unavailable in other locations, thus compromising the ability to monitor often need the resources of an intensive care vital body functions of both mother and baby. Finally, it is worth noting that such results of congestion may negatively affect patients' perceptions of service quality.

Of course, all major capacity decisions need to be made in light of financial constraints, competing demands, and predictions concerning future demands for the service.

Planning for predictable changes in demand

When making capacity decisions about resources that will be used over several years, it is clearly necessary to consider how conditions may change over that period of time. So in determining the choice of arrival rate or average LOS for a queueing analysis of a hospital unit, it would be important to engage in analyses and discussion to gauge how these parameters may change and then run the model to determine the sensitivity of capacity levels to these changes.

Howe the changes in the arrival rate that are
likely to or time-of-year patterns. For example,
obstetrics y in admissions. An analysis performed
on data fr Hospital in Boston (Green and Nguyen
2001) rev low of about 68% in January to about
88% in Ju of the probability of delay of getting a
bed for ar negligible with this capacity. However,
in July, th nd if, as is likely, there are several days
when actu %, this delay probability would shoot up
to over 6 om backups into the labor rooms and
patients o ents. Clearly, hospitals need to plan for
this type capacity that can be used during peak
times, or by using “swing” beds that can be shared by clinical units that have countercyclical demand patterns.



Most hospital units experience different arrival rates for different days of the week. For example, in one surgical intensive care unit, the average admissions per day over a six month period varied from a low of 1.44 for Sundays to a high of 4.40 for Fridays. Using the average arrival rate over the week of 3.34 in a queueing model would indicate that given the 12 bed capacity of this unit, the probability of delay for a bed was about 39%, indicating serious congestion. However, this is very misleading because delays will be significantly greater in the middle of the week and quite small earlier in the week due to the large differences in the admissions rates (Green and Nguyen (2001). This illustrates a situation in which a steady-state queueing model is inappropriate for estimating the magnitude and timing of delays and for which a simulation model will be far more accurate.

Using queueing models to quantify the benefits of flexibility

Healthcare facilities often have to make a choice as to the extent to which resources should be dedicated to specific patient types. For example, should there be a imaging facility just for the use of inpatients, or for emergency patients? Should there be a “fast-track” unit in the emergency room to deal with simpler, non-urgent cases. How many distinct clinical service units should be used for hospital inpatients? In many of these situations, a queueing analysis can be useful in evaluating the potential trade-offs between more flexible and more specialized facilities.

For example, seriously ill patients arriving to a hospital ED often experience serious delays in being admitted due to highly variable patient demands and insufficient inpatient bed capacity. Yet, hospitals are often reluctant or unable to add capacity because of cost pressures, regulatory constraints, or a shortage of appropriate personnel. This makes it extremely important to use existing capacity most efficiently. Increasing bed flexibility can be a key strategy in alleviating congestion. For example, hospitals vary in the degree to which they segregate patients by diagnostic type. While all hospitals have separate units for pediatrics, obstetrics and psychiatric patients, some also have distinct units for clinical services such as cardiology, neurology, oncology, urology, neurosurgery, etc. Other hospitals may make no such distinctions and simply designate all of these as medical/surgical beds. What are the implications of these differing bed assignment policies on delays for beds?

As mentioned in section 2.1, service systems have economies of scale and so in general, the less specialized the beds, the larger the pool of beds that can be used for any type of patient, and therefore the fewer beds should be needed to achieve a given standard of delay. In other words, if one hospital has 100 general medical/surgical beds, and another has the same 100 beds, but allocated into 10 distinct clinical services, each of which can only be used for patients falling into the appropriate category, the second hospital will likely have considerably longer delays for beds (which usually show up as longer stays in the ED) and lower average occupancy levels than the first. This is pretty clear once you consider that by creating separate categories of beds, there is the possibility of patients waiting for beds even when beds are available if they are the "wrong" kind. This also happens when beds are distinguished by capability, for example, telemetry beds.

Clearly, there are many instances in which there are compelling clinical and/or managerial reasons for maintaining particular patient types in specialized units. From a medical perspective, there may be benefits derived from having patients clustered by diagnostic categories in dedicated units managed and staffed by specialized nurses. These include shorter LOS, fewer adverse events and fewer readmits. Yet, many hospital managers believe that nurses can be successfully cross-trained and that increasing bed flexibility is ultimately in the best interests of patients by increasing speedy access to beds and minimizing the number of bed transfers. By incorporating waiting times, percentage of "off-placements" and the effects on LOS, queueing models can be used to better evaluate the benefits of greater versus less specialization of beds or any other resource. This would be done by simply modeling the general-use unit as a single multi-server queueing system fed and comparing the results to those from modeling each distinct service as an independent queue. In the latter case, the overall patient delay can be obtained from an arrival rate weighted average of the individual queue delays (see e.g. Green and Nguyen 2001).

Analyses of Flexible capacity: determining staffing levels to meet time-varying demands

As mentioned previously, health care facilities generally experience very different levels of demand over the day, over the week, and even over the year. Many facilities adjust their staffing – e.g. physicians, nurses, technicians, housekeeping staff – in order to respond to the demands in a timely fashion at minimal cost. This is often done without the help of a quantitative model and can lead to an inefficient and ineffective allocation of resources. Here we use the example of determining physician staffing levels in an ED to illustrate how queueing models can be used to improve performance in these types of situations.

Data collection and model choices

In order to use a queueing model to determine how to adjust staffing to meet time-varying demands, it is first necessary to collect fairly detailed data on the volume of demand that must be handled by that staff by time-of-day and day-of-week. In collecting demand data, the goal is two-fold. First, and most obviously, the data will be used to parameterize the queueing model. However, before that can be done, it must first be determined how many staffing models are needed. That is, will staffing be identical for all days of the week or vary from day to day? For example, in a study conducted in the ED of a mid-size urban hospital in New York City (Green et al 2005), the overall volume varied from a low of 63 patients per day on Saturdays to a high of 72 per day on Monday. This degree of variation indicated that the then-current policy of identical staffing levels for all days of the week was likely suboptimal. However, it was deemed impractical to have a different provider schedule every day and so it was decided to use queueing analyses to develop two schedules: weekday and weekend. This required aggregating ED arrival data into these two groups. For each, demand data was then collected for each hour of the day using the hospital's admissions database to understand the degree of variation over the day (see Figure 5). This level of detail also allows for the use of queuing analysis to determine the impact of different shift starting times on delays and/or staffing levels.

A queueing model also requires an average provider service time per patient, which must include the times of all activities related to a patient. In the ED, these activities include direct patient care, review of x-rays and lab tests, phone calls, charting, and speaking with other providers or consults. In many, if not most, hospitals, these data are not routinely collected. At the time of the study, provider service times were not recorded and had to be estimated indirectly from direct observation and historical productivity data.

Constructing the queueing models

Since the *M/M/s* model assumes that the arrival rate does not change over the day, actual service systems that have time-varying demands typically use this model as part of a *SIPP* (stationary independent period-by-period) approach to determine how to vary staffing to meet changing demand. The *SIPP* approach begins by dividing the workday into staffing periods, e.g. one, two, four or eight hours. Then a series of *M/M/s* models are constructed, one for each staffing period. Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. The service target might be a desired maximum mean delay or probability of delay standard. However, recent research has shown that the *SIPP* approach is often unreliable, particularly when average service times are 30 minutes or more, and that a simple modification, called *Lag SIPP*, is often more effective in identifying staffing levels that achieve the desired performance standard (Green et al 2001). This is because in many service systems with time-varying arrival rates, the time of peak congestion significantly lags the time of the peak in the arrival rate (Green et al 1991). While the standard *SIPP* approach ignores this phenomenon, the *Lag SIPP* method incorporates an estimation of this lag and thus does a better job of identifying staffing levels to limit delays. For the *M/M/s* model, the lag can be well-approximated by an average service time.

Choosing a delay standard and applying the queueing results

In our ED physician staffing study, the *Lag SIPP* approach was applied by first advancing the arrival rate curve by our estimate of the average physician time per patient, 30 minutes. We then constructed a series of *M/M/s* models for each 2-hour staffing interval, using the average arrival rate for each based on the time-advanced curve and the average 30 minute service time. The delay standard we choose was that no more than 20% of patients wait more than one hour before being seen by a provider. The use of one hour is consistent with the time standards associated with emergent and urgent patient groups used in the National Hospital Ambulatory Medical Care Survey (McCaig and Burt 2002). The 20% criterion reflects the approximate percentage of non-urgent arrivals at the study institution.

The modeling results gave the number of ED physicians needed in each of the 2-hour staffing intervals to meet the delay standard. In total, 58 physician-hours were needed on weekdays to achieve the desired service standard, which represented an increase of 3 hours over the existing staffing level of 55 hours. Model runs for the weekend indicated that the target performance standard could be achieved with a total of 53 provider-hours. In both these cases, the queueing analyses suggested that some physician hours should be switched from the middle of the night to much earlier in the day. A more subtle change suggested by the model was that the increase in staffing level to handle the morning surge in demand needed to occur earlier than in the original schedule. Though resource limitations and physician availability prevented the staffing suggested by the queueing analyses from being implemented exactly, the insights gained from these analyses were used to develop new provider schedules. More specifically, as a result of the analyses one physician was moved from the overnight shift to an afternoon shift, 4 hours were moved from the weekends and added to the Monday and Tuesday afternoon shifts (since these were the two busiest days of the week) and a shift that previously started at noon was moved to 10 AM. These changes led to shorter average delays and a reduced fraction of patient that left before being seen by a physician.

USING QUEUEING MODELS TO IMPROVE HEALTHCARE DELIVERY: OPPORTUNITIES AND CHALLENGES

As this chapter has illustrated, service systems are very complex due to both predictable and unpredictable sources of variability in both the demands for service and the time it takes to serve those demands. In healthcare facilities, decisions on how and when to allocate staff, equipment, beds, and other resources in order to minimize delays experienced by patients are often even more difficult than in other service industries due to cost constraints on the one hand and the potentially serious adverse consequences of delays on the other hand. Therefore, it is imperative that these decisions should be as informed as possible and rely upon the best methodologies available to gain insights into the impact of various alternatives.

Queueing theory is a very powerful and very practical tool because queueing models require relatively little data and are simple and fast to use. Because of this simplicity and speed, they can be used to quickly evaluate and compare various alternatives for providing service. Beyond the most basic issue of determining how much capacity is needed to achieve a specified service standard, queueing models can also be useful in gaining insights on the appropriate degree of specialization or flexibility to use in organizing resources, or on the impact of various priority schemes for determining service order among patients.

On the other hand, though queueing models don’t require much data, the type of operational data needed as input to a queueing model is often unavailable in healthcare settings. Specifically, though demand or arrival data are often recorded, service times are usually not documented. So a queueing analysis might require a data collection effort to estimate, for example, the time that a care provider spends with a patient. However, as information technology systems become more prevalent in healthcare, this type of data will be increasingly available.

In developing the data inputs for a model, it’s also very important to make sure that all of the data needed for the model is collected and/or estimated. On the demand side, this means including all demands for care, including the ones that may not have been met in the past because of inadequate capacity. For example, in a hospital ED, some patients who are forced to wait a long time before seeing a physician leave the ED before being seen. If these are not captured in the data collection system that is being used to measure demands, the model will underestimate the capacity needed to meet the desired performance standard. On the service side, it’s important to include all of the time spent by the servers that is directly associated with caring for the patient. For a physician, this may include reviewing medical history and test results in addition to direct examination of the patient.

In addition to data, a queueing analysis of a particular healthcare system requires the identification of one or more delay measures that are most important to service excellence for that facility. These measures should reflect both patient perspectives as well as clinical realities. For example, though hospital ED arrivals with non-urgent problems may not require care within an hour or so from a clinical perspective, clearly very long waits to see a physician will result in high levels of dissatisfaction, and perhaps even departure, which could ultimately lead to lost revenue. Trying to decide on what might be a reasonable delay standard in a specific healthcare facility is not trivial due to a lack of knowledge of both patient expectations as well as the impact of delays on clinical outcomes for most health problems.

In summary, healthcare managers are increasingly aware of the need to use their resources as efficiently as possible in order to continue to assure that their institutions survive and prosper. This is particularly true in light of the growing threat of sudden and severe demand surges due to outbreaks of epidemics such as SARS and avian or swine flu, or terrorist incidents. As this chapter has attempted to demonstrate, effective capacity management is critical to this objective as well as to improving patients' ability to receive the most appropriate care in a timely fashion. Yet, effective capacity management must deal with complexities such as tradeoffs between bed flexibility and quality of care, demands from competing sources and types of patients, time-varying demands, and the often differing perspectives of administrators, physicians, nurses and patients. All of these are chronic and pervasive challenges affecting the ability of hospital managers to control the cost and improve the quality of healthcare delivery. To meet these challenges, managers must be informed by operational and performance data and use these data in models to gain insights that cannot be obtained from experience and intuition alone. Queueing analysis is one of the most practical and effective tools for understanding and aiding decision-making in managing critical resources and should become as widely used in the healthcare community as it is in the other major service sectors.

REFERENCES

Allen, A.O., 1978, *Probability, statistics and queueing theory, with computer science applications*. New York, Academic Press.

Brecher, C. and Speizio, S., 1995*, Privatization and Public Hospitals*, Twentieth Century Fund Press, N.Y.

Brewton, J.P., 1989, Teller staffing models, *Financial Manager's Statemen*t, July-August: 22-24.

Brigandi, A.J., Dargon, D.R., Sheehan, M.J. and Spencer III, T., 1994, AT&T's call processing simulator (CAPS) operational design for inbound call centers, Interfaces 24: 6-28.

Brockmeyer, E., Halstrom, H.L., and Jensen, A., 1948, The life and works of A.K. Erlang, *Transactions of the Danish Academy of Technical Science* 2.

Brusco, M.J., Jacobs, L.W., Bongiorno, R.J., Lyons, D.V. and Tang, B., 1995, Improving personnel scheduling at airline stations, Operations Research, 43: 741-751.

Chelst, K. and Barlach, Z., 1981, Multiple unit dispatches in emergency services, *Management Science*, 27: 1390-1409.

Cobham, A., 1954, Priority assignment in waiting line problems, *Operations Research*, 2: 70-76.

Freeman, R.K., and Poland, R.L., 1997, Guidelines for Perinatal Care, 4th ed., American College of Obstetricians and Gynecologists, Washington, D.C.

Green, L.V., Giulio, J., Green, R., and Soares, J., 2005, Using queueing theory to increase the effectiveness of physician staffing in the emergency department, *Academic Emergency Medicine*, to appear.

Green, L.V., 2003, How many hospital beds? *Inquiry*, 39: 400-412.

Green, L.V., Kolesar, P.J., Svoronos, A., 2001, Improving the SIPP approach for staffing service systems that have cyclic demands, *Operations Research,* 49: 549-564.

Green, L.V. and Nguyen, V., 2001, Strategies for cutting hospital beds: the impact on patient service. *Health Services Research*, 36: 421-442.

Green, L.V., Kolesar, P.J., and Svoronos, A., 1991, Some effects of nonstationarity on multi-server Markovian queueing systems. *Operations Research*, 39: 502-511.

Green, L.V., and Kolesar, P.J., 1984, The feasibility of one-officer patrol in New York City, *Management Science* 20: 964-981.

Hall, R.W., 1990, *Queueing Methods for Service and Manufacturing*. New Jersey: Prentice Hall.

Holloran, T. J. and Byrne, J.E., 1986, United Airlines station manpower planning system, *Interfaces*, 16: 39-50.

Green, L.V., 2003, How many hospital beds? *Inquiry*, 39: 400-412.

Institute of Medicine, Committee on Quality of Health Care in America, 2001, *Crossing the quality chasm: a new health system for the 21st century*. Washington, D.C.: National Academy Press.

Kaplan, E.H., Sprung, C.L., Shmueli, A., and Schneider, D., 1981. A methodology for the analysis of comparability of services and financial impact of closure of obstetrics services. *Medical Care*, 19: 395-409.

Kim, S., Horowitz, I., Young, K.K., and Buckley, T.A., 1999, Analysis of capacity management of the intensive care unit in a hospital, *European Journal of Operational Research* 115: 36-46.

Kolesar, P.J., Rider, K., Crabill, T., and Walker, W., 1975, A queueing linear programming approach to scheduling police cars, *Operations Research*, 23: 1045-1062.

Larson, R.C., 1972, *Urban Police Patrol Analysis*, MIT Press, Cambridge.

McCaig, L.F., and Burt, C.W., 2004, National hospital ambulatory medical care survey: 2002 emergency department summary. *Advance Data from Vital and Health Statistics*, 340: 1-35.

Stern, H.I. and Hersh, M., 1980, Scheduling aircraft cleaning crews, *Transportation Science*, 14: 277-291.

Taylor, P.E. and Huxley, S.J., 1989, A break from tradition for the San Francisco police: patrol officer scheduling using an optimization-based decision support system, Interfaces, 19: 4-24.

Worthington, D.J., 1987, Queueing models for hospital waiting lists. Journal of the *Operations Research Society*, 38: 413-422.

Young, J.P., 1965, Stabilization of inpatient bed occupancy through control of admissions, *Journal of the American Hospital Association*, 39: 41-48.

Table 1 Probability of delay and utilization for obstetrics unit

No. Beds	**Pr(Delay)**	**Utilization**
45	0.666	0.953
46	0.541	0.933
47	0.435	0.913
48	0.346	0.894
49	0.272	0.875
50	0.212	0.858
51	0.163	0.841
52	0.124	0.825
53	0.093	0.809
54	0.069	0.794
55	0.051	0.780
56	0.037	0.766
57	0.026	0.753
58	0.018	0.740
59	0.013	0.727
60	0.009	0.715
61	0.006	0.703
62	0.004	0.692
63	0.003	0.681
64	0.002	0.670
65	0.001	0.660